Elie Naulleau, West Group, USA
Profile-guided terminology extraction

Abstract
This paper presents a noun phrase filtering system designed to retain noun phrases that conform to a certain model. This model is built from data provided by the user and made of samples of phrases that the user would keep or throw away depending on his/her goal. The following motivates this approach: (1) *there is no multi-purpose term extraction grammar*. Even a single document could be considered from multiple points of view, thanks to distinct extraction grammars, distinct sets of term candidates may be extracted from the same document, in order to satisfy different objectives. (2) *The expert/terminologist should participate in the conception of the filtering system*. We acknowledge and integrate the expertise and objectives of the individual who builds the selection system and tunes its behavior. Our technique is to invoke expertise before the filtering process, and to make it reusable, instead of calling its services afterwards (sorting out, manual selection), without being able to reuse it over other documents.

1 Introduction

This work was originally part of a terminology acquisition framework. The tool we describe can be defined as a noun phrase filtering system. Depending on how it is tuned, it can select phrases, expected to be term candidates (for thesaurus updating for instance), or any other kind of phrases (for example, phrases used to create free indexes in documents).

1.1 Semi-controlled noun phrase extraction

The system selects from a list of polylexical noun phrases the ones which morphological, syntactic and semantic descriptions are modeled as relevant in a filtering profile. To define a profile, the user has to sort out among a set of phrases those he considers to be relevant for his work, and those he does not wish to be kept. This sample data is then processed by a earning/generalization procedure that yields a profile. The resulting profile maintains phrases that look like the ones used to define the positive characteristics of the profile, and discards those that look like the ones declared undesirable. In this sense, this filtering process may be described as semi-controlled, and the approach may be seen as a partial and rough capture of the knowledge of the user.

1.2 A filtering process relying on lexico-syntactic dependencies

The selected approach is symbolic (i.e., it does not use any statistical measure). It must be distinguished from other symbolic approaches inasmuch as it does not precipitate and fix the extraction goal in a grammar but learns and stores it in a profile from the specifications of the end-user. We agree with [BOURIGAULT&HABERT98] in that the terminological nature of a phrase cannot be determined out of the scope of a field and even an application. The extraction pattern technique, restricting the structure and the length of the phrases, has not been chosen because it is also a source of noise in large-scale industrial applications [STA97]. This method has been used in terminology acquisition grammars, like the one run by the *AlethIp* engine (Erli) [OGONOWSKY&ALL94; HERVIOU95; HERVIOU-PICARD96]. Moreover, the splitting of maximal noun

phrases[1] with heuristics like those used in the *Lexter* software [BOURIGAULT94] is not used either. These two techniques and their combination, and sometimes their hybridization with statistical measures [DAILLE94; SMADJA93] cover most of the strategy used for the extraction and selection of syntagms, except for purely statistical techniques based on collocations [CHURCH89]. Our strategy relies on the evaluation of the relevance of elementary syntactic dependencies[2] found in the phrases to be filtered. This relevance is always established from the sample data provided by the user of the system.

The way we use elementary lexico-syntactic dependencies is inspired from works derived from distributional analysis [SAGER96; HABERT&AL96]: noun phrases are normalized as lists of dependencies[3] and are evaluated according to the approved combinations between their lexical units. These combinations can be seen as the restrictions of selection peculiar to the point of view defined in the profile.

In this paper we explain how runs the filtering system and we relate two experiments. The first focused on French technical texts and tests the use of two different profiles against the same data. The second was conducted on legal American texts and partially evaluated by a domain expert.

2 Two filtering profiles : technological survey and thesaurus updating

The experiment was carried out at the Direction des Études et Recherches d'Electricité de France (henceforth referred to EDF) [NAULLEAU98]. The aim was to provide tools to update documentary resources, such as the EDF thesaurus[4], to update its coverage for application such as when used in automatic indexing. The corpus was made of internal EDF documents: 1,780 ARDs[5] covering years 1984 to 1995 (about 900,000 words). Although it deals mainly with nuclear equipment, this corpus also deals with numerous connected and intricate fields, making automatic processing difficult.

~~1.1~~2.1 Linguistic enrichment : tools and processes required

Our implementation for the French language is based on tools and resources available at the DER-EDF: the *AlethIP* engine[6] running a grammar producing morpho-syntactic trees of sentences[7]. At the end of this stage, each sentence of the corpus is lemmatized, typed for part-of-speech and syntactically parsed.

These sentences are then enriched with suffix information assigned to nouns and adjectives. These suffixes, taken from Guilbert's work [GUILBERT70], bring raw semantic values (for instance,

---

[1] A syntagm called maximal is usually extracted between predetermined boundaries (verbs, relative pronouns, conjunctions, and certain propositions…) and corresponds to a non-split form.

[2] We define an elementary dependency as binary dependencies (in the sense of [MELĆUK88]) – combined if necessary – expressing a relationship between a head and a modifier (or a predicate and an argument) in their order of occurrence in the statement. For instance from the analysis of the syntagm "*support de ligne électrique aérienne en béton*" we extract the following dependencies: *support→en→béton, support→de→ligne, ligne→électrique, ligne→aérienne*.

[3] which allows us to cast off the yoke of the complexity of processing full syntactic trees.

[4] This thesaurus contains more than 20,000 entries (13,000 descriptors, 7,000 synonyms). It is organized according to 45 general themes, which are subdivided into 330 semantic fields.

[5] An ARD (Actions de Recherche et de Dévelopement) is a short text, written by an EDF researcher, which describes his state of work, activities and goals.

[6] The AlethIP engine is a product of the Erli Company (France).

[7] This grammar is coming from the GRAAL project [SABBAGH&TEAM94], and is operating with the Genelex *AlethDic* 1.1.5 dictionary (Erli). The parser is robust: large amounts of heterogeneous texts can be parsed, often harming the quality of the output. We accepted these imperfections and put the emphasis on ulterior stages.

in French –*aire* in *actionnaire*, *disquaire*, indicates an agent, a function, a job) and can be processed as "fuzzy" semantic categories. In addition, semantic tags are assigned to nouns and some adjectives. The semantic lexicon comes from the semantic layer of the *AlethDic* dictionary[6], simplified from 372 to 72 tags in order to ensure domain independence of any tag and to make the disambiguation task easier. This task is performed thanks to contextual rules [NAULLEAU&AL96] written for the more frequent and ambiguous words in the corpus. Only a few types of adjectives could be tagged, due to the encoding challenge they represent in terms of semantic categorization.
The next stage extracts maximal noun phrases from the syntactic tree output by *AlethIP*'s grammar. A subsequent stage extracts elementary syntactic dependencies from the noun phrase trees.

~~1.2~~2.2 Filtering profiles : a way to formalize the possible from the observed
We have tested two filtering profiles. The first is motivated by a technological survey approach. It has been manually built with the help of an EDF information expert, using a dedicated user interface managing the positive and negative descriptions to be integrated into the profile. The second profile was intended to be used for terminology acquisition in a narrow domain. It has indeed been built from the content of 2 fields of the EDF thesaurus (i.e., "*appareillage mecanique*" –mechanical gear– and "*sciences physiques*" – physic sciences). The table gives examples of entire syntagms or dependencies accepted or rejected for each of the two profiles.

Table 1~~1~~- Positive and ~~negative~~ examples for each profile

Technological survey profile	*Terminology acquisition profile*
usage maritime	zone inondable
~~grande précision~~	~~propagation d'une onde~~
faible hauteur d'eau	~~faible~~ hauteur d'eau
~~le cas d'~~une vallée large	~~codes bidimensionnels d'écoulement~~
calcul des courants	~~calcul des courants~~
besoins de la protection civile	

~~1.1.1~~2.2.1 Learning through generalization and fading of linguistic constraints
A filtering profile is an organized set of linguistic descriptions stored in their complete forms as well as in various intermediate and under-specified forms. It is divided into two subsets: relevant and non-relevant descriptions with respect to the filtering goal. These linguistic descriptions are shaped with the following attributes: lemmatized words, part-of-speech, number, suffix category, semantic tag when available, type of article, type of preposition, dependency relationship between two words. When certain combinations of these items are under-specified (i.e., one or several of them is/are omitted), the profile is granted a predictive power that allows it to consider lexical dependencies that were not used to build it. For instance, omitting the lexical form and using semantic tags allows the usage of hyperonimic abstractions; from "*réacteur à eau*" (water reactor) and "*réacteur à graphite*" (graphite reactor), the program can find a common pattern: *"réacteur à NOM-DE-MATIÈRE*" (SUBSTANCE-NAME reactor). This pattern generalizes the form of the phrases and is able to grab other phrases that match it; for example: "*réacteur à uranium*" (uranium reactor) and "*réacteur à plutonium*" (plutonium reactor).

Table 2~~2~~- Positive sample of modifiers for the name robinetterie

/frNom{Csem=26[8];Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=141}	METALLIQUE
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=142}	AUTOMATIQUE
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=139}	NUCLEAIRE
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=140}	INDUSTRIEL

Table 3~~3~~- Negative sample of modifiers for the name robinetterie

/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=128}	IMPORTANT
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=119}	RECENT
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=130}	NECESSAIRE
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Xcons=6}	EXEMPT
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Xcons=6}	SUJET

~~1.1.2~~2.2.2 Example of prediction on a minimal profile

This mechanism is now illustrated with a more complete example. Tables 2 and 3, respectively, declare relevant and non-relevant lexical dependencies for a given filtering goal, resulting in a profile of 9 dependencies. The profile is built generating the under-specified dependencies from the fully specified ones. The common descriptions between the positive and negative parts are erased because they are not distinctive enough for the task. Without relaxing constraints, the profile would accept and reject the dependencies declared only in tables 2 and 3. Relaxing the constraint on number, the profile would accept the same dependencies, either plural or singular. Tables 4 and 5 show accepted and rejected dependencies replacing the noun and the adjective by their semantic tag, allowing nouns and adjectives of the same semantic tag.

Table 4~~4~~- Accepted dependencies when noun is relaxed but not its semantic tag

Robinetterie métallique,	turbine métallique,	pompe métallique, …
Robinetterie automatique,	turbine automatique,	pompe automatique, …
Robinetterie nucléaire,	turbine nucléaire,	pompe nucléaire, …
Robinetterie industrielle,	turbine industrielle,	pompe industrielle, …

Table 5~~5~~- Accepted and rejected(x) dependencies when adjective is relaxed but not its semantic tag

robinetterie métallique,	robinetterie inoxydable,	robinetterie poreuse
robinetterie automatique,	robinetterie isotherme,	robinetterie modulaire
robinetterie nucléaire,	robinetterie électromagnétique,	robinetterie radioactive
robinetterie industrielle,	robinetterie communautaire.	
xrobinetterie importante,	xrobinetterie conventionelle,	xrobinetterie particulière
xrobinetterie sujette (à),	xrobinetterie relative (à),	xrobinetterie envisagée (par)
xrobinetterie récente,	xrobinetterie ancienne	
xrobinetterie nécessaire,	xrobinetterie indispensable	

[8] Meaning of the feature values:

Csem = 26	Device noun (ENTITÉ-CONCRET-ARTEFACT-APPAREIL)
Morpho = 2	feminine singular
Csem = 119	temporal localization in the past
Csem = 128	qualifying subjectively
Csem = 130	deontic value
Csem = 139,142	various properties (incomplete coding)
Csem = 141	relationship with a substance
Csem = 142	« relative to » the derived noun (ex : industrial : relative to industry)
Xcons = 6	Adjectives built with a preposition (*exempt de, nécessaire à, sujet à*)

~~1.3~~2.3 A filtering process pruning undesirable dependencies

The filtering process consists in projecting the syntactic dependencies of the profile and their status (relevant/non relevant) on the nominal phrase trees parsed by *AlethIP*. The search for the dependencies in the profile proceeds from the more specified descriptions to the less specified ones. The non-relevant dependencies are removed from the tree. For example, on the noun phrase: *« amélioration de la connaissance des phénomènes dans la zone d'assèchement des tubes de GV chauffés au sodium »*, given the fact that *amélioration→de→connaissance, connaissance→de→phénomène,* and *phénomène→dans→zone* are found non relevant in the profile, the final pruned phrase is *«zone d'assèchement des tubes de GV chauffés au sodium »*. In other cases, several sub-trees can be found at the end of this stage. For instance, the tree of the phrase *«programme expérimental de validation de code de calcul d'écoulement diphasique dans les faisceaux de tube»* is split into two final phrases: *«code de calcul d'écoulement diphasique* and *«faisceaux de tube»* for the technological survey profile. The two profiles have been run against a sub-collection of the corpus, the 1994 ARDs, in which the *AlethIP* grammar identified 10,936 maximal noun phrases. After filtering, 8,526 noun phrases have been kept by the technological survey profile (77.9%), and 999 phrases have been retained by the terminology acquisition profile (9.1%), 420 phrases being shared by the two profiles.

2.4 Evaluation : limited but useful semantic tagging

The results could not be manually evaluated because of the scarcity of experts[9]. However, we have tested several learning modes, varying the size of the learning sample or the constraints used to build the profiles. We found that the performance of the system (ability to identify a relevant dependency) falls by 20% when semantic tags are not used to build the profile. In addition, when size of the learning sample varies from 90% to 60% of the available descriptions, the resulting profile shows performance degradation of 10%. We can assume here that the generalization based on the linguistic attribute gives a quite efficient prediction: with 60% of the available descriptions, the system almost reaches the level of performance it has with 90% of the available descriptions.

3 Experiment on American legal texts

West Group[10] provides value-added legal information to its customers. Statutes, regulations, case law, and many other kinds of documents are indexed, annotated, enriched and clarified by human editors. We have performed an experiment on judicial opinions (i.e., legal cases). These texts represent the American system of jurisprudence. They are full of facts, details on specific issues but also contain high level concepts, such as legal notions and decisions.

~~1.1~~3.1 Building semi-automatically a filtering profile for legal term identification

The human editor adds headnotes to the cases. Headnotes are intended to outline or clarify points of law. They are written in the sub-language of the editors. Our goal is to identify legal terms[11] in

[9] Maintaining or increasing the availability of experts who could validate the results has not followed from the industrialization of terminology acquisition. The evaluation of the contribution of such tools is still difficult to set up.

[10] West Group is a division of The Thomson Corporation and the leading provider of information to the U.S legal market.

[11] It is difficult to speak about a legal terminology in the context of information retrieval: any term from daily life can be brought into a legal context and be used for legal purposes. However there are terms particularly legal when we consider their origin and usage. We are trying to identify phrases that conform to the sub-language of the editors and that have an indexable content.

the cases, before they are enriched with headnotes. This is not a terminology acquisition task but a noun phrase-filtering task. The selected phrases are then used in document categorization tasks.

~~1.1.1~~3.1.1 A profile pulled out from documents

The approach for the English language tends to be more approximate and less time consumptive than for the French experiment. For this reason, we tried to build a first profile without manual intervention. One can note that points of law and cases contain legal terms. However these terms appear more systematically in points of law, without any other kind of material. We have selected an 18 million words corpus of headnotes. We have used it to build the positive part of the profile. We have also selected a corpus of cases of 14 million words (without their headnotes) to build the negative part of the profile. The profile has been built by extracting the maximal noun phrases and their dependencies from the headnotes and the cases. The intersection of these two sets of dependencies (29% of common material) has been discarded to make the positive and negative parts more discriminating (these dependencies will be considered neutral while filtering).

Table ~~6~~6 - Content and size of legal term profile

Descriptions	*Number of complete dependencies*	*Number of under-specified dependencies*
Positive	145 717	271 834
Negative	109 874	210 642
Examples[12]	*mod*:presumptions→of→proof	*mod*:NNS→of→proof *mod*:presumptions→of→NN
	attr:probable→cause	*attr*:JJ→cause, attr:probable→NN
	attr:good→faith	*attr*:JJ→faith, attr:good→NN

3.1.2 Some robust but still perfectible linguistic processing

The first stage of the process is a part-of-speech tagging performed by TreeTagger [SCHIMD94]. A dedicated tool we developed then performs a noun phrase extraction. It is based on a combination of the technique of repeated segments [SALEM87] and the use of the part-of-speech information. The lexico-syntactic dependencies are generated with about 1,500 pattern-based rules unfolding dependencies from the original phrase. Such a technique lacks accuracy compared to the output of a parser. We are considering using S. Sekine's parser [SEKINE98] to improve parsing accuracy. Furthermore, no semantic tagging has been done. As a result, the profile will perform more radically, inducing generalizations on the sole part-of-speech. According to the results on French data, one can expect a performance loss of 20%. We have relied on the large size of the training data to moderate the somewhat wild generalization. Table 6 shows the content of the profile in terms of the number of complete and under-specified dependencies.

3.2 A simplified filtering without syntactic pruning

The data submitted for noun phrase extraction and filtering was made up of 7 million words collected from cases. The extraction brought back 75,920 noun phrases[13]. Unlike what has been done for French, the filtering does not alter the structure of the noun phrases but assigns a relevance score to them. The score is established by counting the negative and positive

[12] Read NN for 'noun', NNS for 'plural noun', JJ for 'adjective'; *mod* tells about a head/modifier or predicate/argument relationship; *attr* means an attributive relation.

[13] Each of these phrases can occur from two to several thousand times.

dependencies in the phrases, according to the profile. Unknown dependencies are also taken into account. A weaker confidence is assign to under-specified dependencies.

Table 7: Filtering legal terms – sample of results

Excluded phrases	*Retained phrases*
collateral consequences, clear weight, clear distance, class D, civil rules, challenged ordinances, business privilege, base solutions, arthroscopic surgery, arbitrary factor, lack of personal involvement, malicious destruction, household member, principal places of business, MDI's alleged trade dress, Northern District of Florida, meal and beverage subsidy, dangers of cigarette smoking, evidence of sex discrimination, same-sex harassment	intentional infliction of emotional distress, United States District Court, equal protection, burden of proof, deposition testimony, judgment of the trial court, law claims, sexual harassment, sex harassment, law enforcement, habeas corpus, findings of fact, judgment of the circuit court, criminal proceedings, contractual obligations, doctrine of qualified immunity, denied effective assistance of counsel, judgment of the court of appeals

3.3 Results and manual evaluation by a domain expert

After filtering, 28,556 phrases are retained. 20% of them are kept thanks to under-specified description in the profile. 44,699 are thrown away, 35% of them thanks to under-specified descriptions. There are 2,665 which have not been classified, due to the lack of information in the profile. For the evaluation, 1,000 noun phrases selected by random (500 kept, 500 discarded) have been submitted to a legal expert. According to him, 58% of the phrases have been well categorized (legal/non legal term). This result is quite weak. The absence of semantic tagging is probably one of the causes. But it turns out that the most significant mistakes occur where the profile has been neutral, i.e. where 29% of the common dependencies had been neutralized. So to refine this profile, the expert will have to decide where to put the phrases that yielded the common dependencies, either in the negative or positive side of the profile.

4 Conclusion

Our system has the following characteristics:

i. It must be calibrated for a given task or domain
ii. It requires that the user (terminologist, expert) actively participates in the process of defining the filters, and providing to the system with learning samples.
iii. The evaluation of the relevance of any phrase relies on a lexico-semantic description of its syntactic dependencies.

The experiment performed on French shows that there is an actual benefit from using semantic tags, even rough tags. The experiment on English shows that one can initiate the definition of profile from documents, when their content matches the goal of the filtering.

A profile tries to predict the acceptability of new syntactic dependencies from the description of already observed dependencies. In this context, the profile takes advantage of two distinct and complementary aspects of the competence of the expert: what s/he retains, what s/he discards. In this sense a profile attempts to capture the expertise of an individual or of a group of individuals. As a result their practical experience is summed up and stabilized. The dependencies play a core role in the filtering process. Beyond this, they allow one to bypass the complexity of full

syntactic processing; they are lexico-syntactic handles that bridge the gap between single words and phrases, highlighting head–modifer/argument relationships. There are however some cases, when they fail to distinguish what must be kept or discarded. For instance if we had to throw away "*30 days*" but to keep "*30 days warranty*", the common dependency *30→days* would create a conflict when deciding to put it in the negative or positive side of the profile. For now, it is made neutral, i.e. simply removed from the profile. One solution is to take into account the tree-like contexts of the dependency, which is a way to induce a form of grammar, as shown in [GAUSSIER&HABERT97].

Future work

Filtering verbal phrases is not out of the scope of our system. It would require submitting to the system dependencies extracted from the verb constellation. In so far as the parsing would generate correct dependencies for any word of a sentence, single words could also be filtered the same way.

A larger evaluation could tell about the actual efficiency of the dependencies for noun phrase filtering. There is two ways to do this: the manual one – the user/author of a profile would have to validate the results on large amount of data – and the automatic one – one could try to reproduce the results of another filtering system considering its output as training data for our system.

Acknowledgements

The experiment of French was part of the author's thesis which was sponsored by a CIFRE grant, funded jointly by the Association Nationale pour la Recherche technique, the École Normale Supérieure de Fontenay/St Cloud and the Direction des Études et Recherches d'Electricité de France. I thank Benoit Habert (ENS Fontenay) for his support and review and also Richard Quatrain (EDF-DER) for his information expertise. I thank Tom Curran (West Group) for providing his legal expertise. I thank Isabelle Moulinier (West Group) and Jack Conrad (West Group) for their review of the English version of this paper.

References

BOURIGAULT D. (1994). LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes. Thèse en mathématiques, informatique appliquée aux sciences de l'homme, École des Hautes Études en Sciences Sociales, Paris.

BOURIGAULT D. & HABERT B. (1998). Evaluation of terminology extractors: principles and experiments. In A. RUBIO, N. GALLARDO, R. CASTRO & A. TEJADA, Eds., First International Conference on Language Resources and Evaluation, volume I, p. 299-305, Granada.

CHURCH K. W. (1989). A stochastic parts program noun phrase parser for unrestricted text. In Proc. ICASSP, p. 695-698, Glasgow (Scotland).

DAILLE B. (1994). Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques. Thèse en informatique fondamentale, Université de Paris 7, Paris.

GAUSSIER E. & HABERT B. (1997). Langue spécialisée : des séquences observées aux mots possibles. In D. CORBIN, B. FRADIN, B. HABERT, F. COISE KERLEROUX & M. PLÉNAT, Eds., Mots possibles et mots existants, Lille.

GUILBERT L. (1970). Fondements lexicologiques du dictionnaire - de la formation des unités lexicales. Grand Larousse de la Langue Française.

HABERT B., NAULLEAU E. & NAZARENKO A. (1996). Symbolic word clustering for medium-size corpora. In 16th International Conference on Computational Linguistics, volume 1, p. 490-495, Copenhague, Danemark.

HERVIOU M.-L. (1995). Applications d'extraction des connaissances à EDF-DER. In Actes de IA'95, Montpellier.

HERVIOU-PICARD M.-L. (1996). Les outils d'indexation AlethIP issus du projet GRAAL : principes et utilisation. Rapport interne HN-46/96/022, EDF, Direction des Études et Recherches, Clamart.

LIN D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In Proceedings of ACL-97, Madrid.

MELČUK I. (1988). Dependancy Syntax. New York: SUNNY NY.

NAULLEAU E. (1998). Apprentissage et filtrage syntaxico-sémantique de syntagmes nominaux pertinents pour la recherche documentaire. PhD thesis, Université Paris 13. Dowloadable at http://www.ens-fcl.fr/~ naulleau/download.html.

NAULLEAU E., HABERT B. & MONTEIL M.-G. (1996). Tagging term components with semantic information. In C. GALINSKI & K.-D. SCHMITZ, Eds., Terminology and Knowledge Engineering, p. 110-117, Vienna: INDEKS-Verlag.

OGONOWSKI A., HERVIOU M. & DAUPHIN E. (1994). Tools for extracting and structuring knowledge from texts. In ACOL, p. 1049.

SABBAGH S. & TEAM G. (1994). Graal eureka project : re-usable grammars for automatic language analysis. In Proceedings of the Language Engineering Convention, Paris.

SAGER N. (1987). Computer processing of narrative information. In N. SAGER, C. FRIEDMAN & M. S. LYMAN, Eds., Medical Language Processing: Computer Management of Narrative Data, chapter 1, p. 3-22. Addison-Wesley.

SALEM A. (1987). Pratique des segments repétés: essai de statistique textuelle. Paris: Kliencksiek.

SCHMID H. (1994). Probalistic part-of-speech tagging using decision trees. In Proceedings of NMLP'94, Manchester, UK.

SEKINE S. (1998). Corpus-based Parsing and Sublanguage Studies. PhD thesis, New York University.

SMADJA F. (1993). Retrieving collocations from text: Xtract. Computational Linguistics, 19(1), 143-177. Special Issue on Using Large Corpora: I.

STA J.-D. (1997). Acquisition Terminologique en Corpus : aspects linguistiques et statistiques. PhD thesis, Université Paris 7, Paris.